Estimation of a Predictor's Importance by Random Forests When There Is Missing Data : RISK Prediction in Liver Surgery using Laboratory Data

Hapfelmeier, Alexander ; Hothorn, Torsten ; Riediger, Carina ; Ulm, Kurt

DOI: https://doi.org/10.1515/ijb-2013-0038

Research Article

Alexander Hapfelmeier*, Torsten Hothorn, Carina Riediger and Kurt Ulm

Estimation of a Predictor's Importance by Random Forests When There Is Missing Data: RISK Prediction in Liver Surgery using Laboratory Data

Abstract: In the last few decades, new developments in liver surgery have led to an expanded applicability and an improved safety. However, liver surgery is still associated with postoperative morbidity and mortality, especially in extended resections. We analyzed a large liver surgery database to investigate whether laboratory parameters like *haemoglobin*, *leucocytes*, *bilirubin*, *haematocrit* and *lactate* might be relevant preoperative predictors. It is not uncommon to observe missing values in such data. This also holds for many other data sources and research fields. For analysis, one can make use of imputation methods or approaches that are able to deal with missing values in the predictor variables. A representative of the latter are Random Forests which also provide variable importance measures to assess a variable's relevance for prediction. Applied to the liver surgery data, we observed divergent results for the laboratory parameters, depending on the method used to cope with missing values. We therefore performed an extensive simulation study to investigate the properties of each approach. Findings and recommendations: Complete case analysis should not be used as it distorts the relevance of completely observed variables in an undesirable way. The estimation of a variable's importance by a self-contained measure that can deal with missing values appropriately reflects the decreased relevance of variables with missing values. It can therefore be used to obtain insight into Random Forests which are commonly fit without preprocessing of missing values in the data. By contrast, multiple imputation allows for the assessment of a variable's relevance one would potentially observe in complete-data situations, if imputation performs well. For the laboratory data, *lactate* and *bilirubin* seem to be associated with the risk of liver failure and postoperative complications. These relations should be investigated by future studies in more detail. However, it is important to carefully consider the method used for analysis when there are missing values in the predictor variables.

Keywords: Random Forests, variable importance, missing data, imputation, liver surgery

DOI 10.1515/ijb-2013-0038

1 Introduction

Due to improvements in liver surgery, postoperative mortality and morbidity decreased dramatically over the last decades [1–3]. However, operations are still not hazard-free [1]. Together with a simultaneously growing application, improved safety is of major concern [4, 5]. For this purpose, the identification of potential risk factors may help to firstly, deepen knowledge about involved processes and secondly, to develop models for the prediction of individual patient outcomes [6–8]. Based on data taken from several

***Corresponding author: Alexander Hapfelmeier,** Institute of Medical Statistics and Epidemiology, Technische Universität, München, Germany, E-mail: Alexander.Hapfelmeier@tum.de
Torsten Hothorn, Division of Biostatistics, Universität Zürich, Zürich, Switzerland, E-mail: Torsten.Hothorn@uzh.ch
Carina Riediger, Department of Surgery, Technische Universität Dresden, Dresden, Germany, E-mail: carina.riediger@tum.de
Kurt Ulm, Institute of Medical Statistics and Epidemiology, Technische Universität, München, Germany, E-mail: kurt.ulm@tum.de

sources like clinical trials, databases, observational studies and many others, one is able to explore these research questions. However, it is important to note that such data, whether they emerge from this research field or any other, often contain missing values. These need to be taken into account by any analysis method as they might have a strong impact on analysis results.

For predictive modeling of the liver surgery data, we used Random Forests [9] which are well known for their ability to deal with missing data. In recent investigations, models fit with and without imputation of missing values showed only negligible differences in predictive performance [the latter was assessed by the mean squared error (MSE)] that Random Forests produced for independent test data [10, 11].

Further strong advantages over common approaches like linear and logistic regression is the ability of Random Forests to implicitly deal with high dimensional data, complex interactions and collinearity (cf. Cutler et al. [12] and Lunetta et al. [13], for corresponding discussions).

In addition, Random Forests provide variable importance measures which can be used to assess a variable's relevance for prediction. In a subsequent step these measures may even be used for variable selection [14–22]. However, the computation of importance measures is not straightforward when there are missing values in the predictor variables. Therefore, a solution to this issue has been proposed in an earlier work [23] (see also Section 2.4). This self-contained measure is closely related to existing approaches and retains appreciated properties, yet handles missing values in an intuitive way without the need for any preprocessing of the data.

Alternative ways to deal with missing values are complete case analysis and imputation (e.g. mean, hot-deck, conditional mean and predictive distribution substitution). However, it has been shown that these ad hoc methods may lead to biased inference when the data are not missing completely at random [24, 25]. Multiple imputation by chained equations (MICE) [26, 27] is meant to solve this problem, and its superiority has been shown in many publications [28, 29].

To our knowledge, it has so far not been investigated how to proceed for the estimation of importance measures when there are missing values in the predictor variables. Therefore, and beyond our previous works, the following analyses investigate the properties of three competing methods, i.e. complete case analysis, multiple imputation and the self-contained importance measure, when used to estimate a variable's relevance for prediction.

In case of the classification problem given by the liver surgery data, preoperative laboratory parameters that contain missing values are used for the prediction of postoperative complications. We found that resulting estimates of a variable's importance depend on the application of complete case analysis, multiple imputation (executed by MICE) and the self-contained importance measure. For a more objective assessment, extensive simulation studies that involve various missing data generating processes were conducted for both, regression and classification problems.

2 Methods

2.1 Missing data

In early works Rubin [30, 31] specifies the issue of correct statistical inference with missing values by the definition of missing data generating processes:

- Missing completely at random (MCAR): $P(R|X_{\text{comp}}) = P(R)$
- Missing at random (MAR): $P(R|X_{\text{comp}}) = P(R|X_{\text{obs}})$
- Missing not at random (MNAR): $P(R|X_{\text{comp}}) = P(R|X_{\text{obs}}, X_{\text{mis}})$

Whether a value is missing is indicated by a binary variable R and depends on its probability distribution $P(R)$. The complete variable set X_{comp} consists of the observed values X_{obs} and the missing ones X_{mis}; $X_{\text{comp}} = \{X_{\text{obs}}, X_{\text{mis}}\}$. Therefore in a MCAR scheme the probability for a missing value is independent of the

observed and unobserved data. By contrast for MAR this probability is dependent on the observed information. In MNAR the probability depends on unobserved variables or the missing values themselves.

Little and Rubin [32] showed that usual sample estimates, for example in linear regression, stay unaffected by the MCAR scheme. By contrast, in classification and regression trees even MCAR may induce a systematic bias, which may be carried forward to Random Forests based on biased split selections [33]. Also, it is well known that complete case analysis is prone to biased inference when the data are not MCAR. Therefore, in the following simulation study, one MCAR, four MAR and one MNAR scheme to generate missing values are investigated.

2.2 Multivariate imputation by chained equations

Single imputation can lead to severe underestimation of variance [34]. A simple and popular solution to this problem is the application of multiple imputation (MI) [31, 35]. In a first step a proper MI approach is supposed to draw M estimates $\theta^{(1)}, \ldots, \theta^{(M)}$ from $P(\theta|X_{\text{obs}})$ for the multi-dimensional parameter θ which determines the data distribution. These are subsequently used in the conditional distributions $P(X_{\text{mis}}|X_{\text{obs}}; \widehat{\theta}^{(t)})$, $t = 1, \ldots, M$ to draw multiple imputations for missing values. This way several imputed datasets are created. Finally, any measure of interest can be assessed by the average of estimates for each of the imputed datasets. Little and Rubin [32] point out that the approach makes standard complete-data methods applicable to incomplete data (e.g. the original permutation importance measure).

The case of more than one variable with missing values demands for a special imputation procedure. A practical approach which makes it possible to bypass the specification of a joint distribution is MICE (also known as imputation by fully conditional specification) [26, 27, 36, 37]. It cycles through incomplete variables to iteratively update imputed values and parameter estimates until convergence. The procedure is repeated several times to produce multiple imputed datasets. An apparent advantage is that imputation of the data can be achieved by a flexible specification of predictive models for each variable.

MICE is especially suitable in MAR settings though Janssen et al. [28] state that it should also be preferred to ad hoc methods like complete case analysis even in MNAR situations. Likewise He et al. [38] and White et al. [27] point out that MICE is also capable to deal with MNAR schemes as the imputation model becomes more general and includes more variables to make MAR plausible.

2.3 Random Forests

The most famous representative of recursive partitioning is the CART algorithm [39]. It constructs trees by sequential binary splits that produce subsets of the data which are as homogeneous as possible in terms of the outcome. Breiman [40] also showed that the performance of single trees benefits from "bagging" (bootstrap aggregation). In bagging, several trees are fit to bootstrapped data. As a further advancement, Random Forests [9, 41] have been introduced for which splits are performed in random selections of variables. This makes a more diverse set of variables contribute to the joint prediction. The latter is found by averaged values or majority votes of predictions given by the trees in a Random Forest. The so-called "out of bag" (OOB) samples, i.e. observations not used to fit the respective trees, can be used for an unbiased estimate of a Random Forests error, viz. the OOB-error.

When there are missing values, surrogate splits need to be employed. They mimic the initial split of the data as they try to achieve the same partitioning of complete observations. When several surrogate splits are computed they can be ranked according to their ability to resemble the initial split. An observation that contains more than a single missing value is processed along this ranking until a decision is found.

The CART and the C4.5 algorithms, and consequently all Random Forest algorithms based on the same construction principles, have been shown to be prone to biased variable selection [33, 39, 42–45]. Therefore, Random Forests used in this work base on the recursive partitioning approach of Hothorn et al. [43]. It

follows the same rationale as Breiman's original approach and guarantees unbiased variable selection and variable importance measures when combined with subsampling [46].

2.4 A variable importance measure for data with missing values

The most popular and most advanced variable importance measure for Random Forests is the permutation accuracy importance measure. It is assessed by the comparison of a trees prediction accuracy, i.e. the rate of correctly classified observations in a classification problem, before and after the random permutation of a predictor variable. If the latter is of relevance for prediction, the accuracy is supposed to drop as the original association with the response and further predictors is destroyed by permutation. The importance measure takes large values in such a case. A decrease of prediction accuracy can also be expressed as an increase of its complement, the prediction error (e.g. MSE). This is a more general formulation as it allows for the computation of the permutation importance measure in regression problems, too.

The variable importance measure estimates a variable's relevance for prediction in a Random Forest. This differs from the estimation of the marginal (or conditional) relation between the variable and the outcome. For example, when the true marginal (conditional) relation is high, the relevance for prediction can still be low. It might just be the case that a Random Forest can simply make no use of a variable, e.g. as it contains many missing values or for any other reason. If the purpose of the importance measure was to estimate the marginal (conditional) relation to the outcome, a low estimate would demonstrate bias in such a case. However, the importance measure is supposed to estimate something else: the relevance of a variable for prediction in a Random Forest. Therefore it is no "bias" but rather a desirable property of an estimator to reflect the decreased relevance of the variable in a forest.

Another issue is that there is no straightforward way to compute the permutation importance measure when there are missing values. In particular, it is not clear how conclusions about a variable's relevance for prediction can be drawn from the permutation approach when surrogate splits, and therefore surrogate variables, are involved in the computation of the prediction error. An alternative approach was proposed earlier [23] to overcome this pitfall. In order to retain appreciated properties it is closely related to existing methodology, yet differs in one substantial aspect: Instead of permuting the values of a variable X_j (that may be missing), observations are randomly sent to the daughter nodes if a parent node k is split in X_j. The probability to be sent left is determined by the relative frequency $\hat{p}_k$ of observations that initially went this way. The algorithm to compute the importance measure is given by:

1. Compute the OOB prediction error of a tree.
2. Randomly assign each observation with $\hat{p}_k$ to the child nodes of a node k that uses X_j for its primary split.
3. Recompute the OOB prediction error of the tree, following step 2.
4. Compute the difference between the original and recomputed OOB prediction errors.
5. Repeat steps 1–4 for each tree and use the average difference of OOB prediction errors over all trees as the overall estimate.

This procedure simulates, like the random permutation in the original permutation importance measure, the null hypothesis that the allocation of observations does not depend on the particular predictor variable (Ishwaran [47] proposes a similar method as alternative to the permutation of variables). It solves any problems associated with the occurrence of missing values and the application of surrogate splits as decisions are detached from the raw values of a variable. No data preprocessing, e.g. imputation is need for this self-contained measure.

In general, Random Forests are highly appreciated for their ability to implicitly deal with missing values in the predictor variables. That is why they are commonly used without any considerations about alternative ways to preprocess missing values, e.g. using imputation. However, the Random Forests approach is often called a "black box" and applicants usually claim for more insight into their prediction

models. The method described above serves this purpose as it closes a gap that existed for the computation of importance measures in such situations.

An alternative and very basic approach to compute a variable's importance in the presence of missing values is to simply count the number of times it is chosen for splits in the numerous trees of a forest. However, it is easy to see that this procedure is not able to properly determine a variable's relevance for prediction, e.g. as it ignores a variable's position in the trees. For example, an important variable that is chosen close to the root nodes of trees might easily be outnumbered by a less important variable that is found close to the leaves of the trees. Furthermore, even useless variables are present in forests as the algorithm has to choose the split criterion from random sets of variables. As a consequence, they are assigned a certain importance >0, i.e. their selection frequency. That is why a simple count of variable selection frequencies will not be used as importance measure in the following. Likewise, the original permutation importance measure will also not be used as it has been shown to produce similar results to the approach presented above when there are no missing values [23].

3 Liver surgery data

The liver surgery data have been prospectively collected between July 2007 and July 2012 in an electronical database in the department of surgery of the Klinik rechts der Isar, Technische Universität München. The database contains clinical information and laboratory tests recorded in 562 liver resections. Of those, 332 observations were suitable for analysis. Laboratory blood tests were analyzed as potential risk factors for postoperative incidents such as mortality, re-operation and morbidity. For a global risk assessment, the latter have been combined in a binary outcome which takes the values "any incident" and "no incident". The predictors *platelets* (#/nl), *haemoglobin* (g/dl), *leucocytes* (G/L), *bilirubin* (mg/dl), *haematocrit* (%) and *lactate* (mg/dl) contained 83 (25%), 80 (24%), 80 (24%), 99 (30%), 194 (58%) and 187 (56%), respectively, missing values. The correlation between these variables is given in Table 1. It was computed for all pairwise complete observations. Most of the correlations are weak to moderate. However, there is a strong correlation between *haemoglobin* and *haematocrit*, $r = 0.975$.

Table 1 Bravais-Pearson correlation of predictors in the liver surgery data

	Haemoglobin	*Leucocytes*	*Bilirubin*	*Haematocrit*	*Lactate*
Platelets	0.248	0.526	−0.111	0.352	−0.023
Haemoglobin		0.308	0.061	0.975	−0.270
Leucocytes			0.020	0.423	0.007
Bilirubin				0.034	0.120
Haematocrit					−0.348

Random Forests of size *ntree* = 500 were fit 100 times to the data. This way, location (quantiles) and scale (variability) of resulting importance measures could be compared between approaches. As a representative of more commonly used methods, logistic regression was applied, too. It was fit to 50,000 samples of the data, each made up by 63.2% of observations. This way, the same number of logistic regression models and trees are fit to data produced by the same sampling method. Stepwise backward variable selection based on the AIC criterion [48] was used as an alternative means to assess a variable's relevance for prediction by logistic regression models. The latter is simply given by the number of times a variable is chosen within the repeated computations [49–51]. Only the complete case analysis and multiple imputation approach could be used to produce results for the logistic regression model.

Analysis results

Figure 1 shows that all approaches agree on *lactate* to be the most relevant predictor for complications after liver surgery. This finding seems to be based on quite strong evidence as *lactate* is given the highest importance although it contains 56% missing values. However, the mean rankings and the variability of values show that there is still some uncertainty and disagreement. Concerning the other variables, there are even more pronounced discrepancies. Firstly, the relations among estimates (i.e. the estimated importance of variables in reference to the most important one) vary a lot, e.g. *bilirubin*. Secondly, there is a concerning discordance in mean rankings which even culminates in totally different ratings for *haematocrit*. It is assigned the third, fourth and second rank using imputation, complete case analysis and the self-contained importance measure, respectively. Equal diversities can be found for the other variables, too. The estimated importances of *platelets* and *haemoglobin* are downgraded when complete case analysis is applied. These values even become negative which implies redundancy [47, 52]. Such a rating is questionable considering the clear positive assessment given by the other methods.

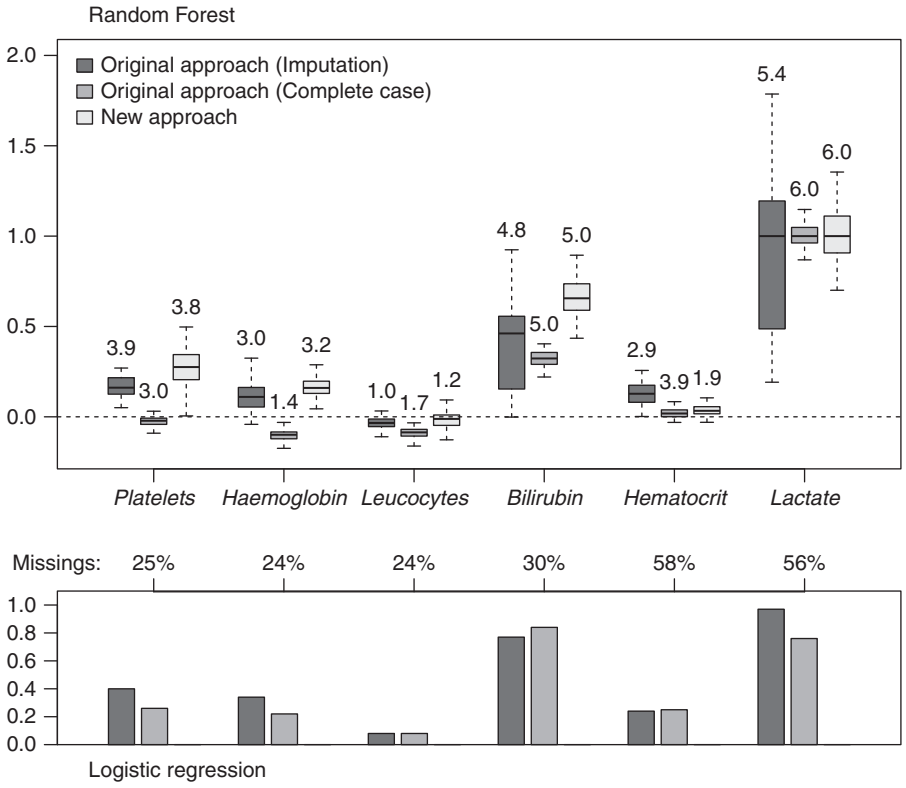


Figure 1 Variable importance assessed for the liver surgery data (application study). Upper: Random Forest. Boxplots display the distribution of importance measures. For a fair comparison values are standardized to the highest median importance within each method. Corresponding mean rankings are given above bars. Most relevant variables are ranked highest. Lower: logistic regression. Bars indicate selection frequencies of variables

Concerning the analysis of imputed data, logistic regression with variable selection produces similar results to the Random Forests variable importance measure. However, using complete case analysis, there are remarkable differences. *Haemoglobin* is now rated to be of higher importance than *leucocytes*. Even more striking, *lactate* that was assigned the highest importance is now ranked after *bilirubin*. In addition to this difference between prediction models, within logistic regression, the results of complete case analysis again differ from those of multiple imputation. After all, it is not surprising to obtain different results from such unequal approaches like Random Forests and logistic regression. As long as the latter is not explicitly

modeled in a very flexible way (i.e. inclusion of interaction terms, non-linear relations, etc.), the former is more capable to deal with (and to reflect) complex (cor-)relations among variables.

To be fair, it has to be mentioned that it is hard to judge the performance of each method in this real data setting. The actual relations of each variable to the outcome are simply unknown. Therefore further simulation studies were set up to investigate the properties of each method in a known setting.

4 Simulation study

An extensive simulation study was designed to investigate the properties of complete case analysis, multiple imputation by MICE and the self-contained importance measure when used for the estimation of a variable's relevance for prediction. In addition, the prediction error of Random Forests that based on each of these approaches was computed for independent test datasets. There are several factors of potential influence that needed to be explored; therefore the amount of missing values, correlation schemes, variable strength and different processes to generate missing values were of particular interest. A detailed explanation of the setup is given in the following.

– *Influence of predictor variables*

The simulated data contained both, a classification and a regression problems. Therefore, a categorical (binary) and a continuous response were created in dependence of six variables with coefficients β:

$$\beta = (1, 1, 0, 0, 1, 0)^\top.$$

Repeated values for β make it possible to compare variables that have, by construction, the same conditional relation to the outcome. The marginal relation, however, is supposed to differ due to the correlation between variables (details are given below, cf. "Correlation"). Another important factor which induces differences between variables is the introduction of missing values. Furthermore, the non-influential variables with $\beta = 0$ help to investigate possible undesired effects and serve as a baseline. An additional analysis with $\beta_0 = (0, 0, 0, 0, 0, 0)^\top$ was performed for a more elaborate investigation of this issue (henceforth called "null case").

– *Data generating models*

A continuous response was modeled by means of a linear model:

$$y = x^\top \beta + \varepsilon \text{ with } \varepsilon \sim N(0, 0.5).$$

The binary response was drawn from a Bernoulli distribution $B(1, \pi)$ with parameter π which was assessed by means of a logistic model

$$\pi = P(Y = 1 | X = x) = \frac{e^{x^\top \beta}}{1 + e^{x^\top \beta}}.$$

The variable set X itself contained $n = 100$ observations drawn from a multivariate normal distribution with mean vector $\mu = 0$ and covariance matrix Σ. To investigate the quality of the data simulation step, linear and logistic regression models were fit to the simulated data. A comparison of the estimated coefficients to β showed low bias and good coverage (cf. Appendix C).

– *Correlation*

$$\Sigma = \begin{pmatrix} 1 & 0.3 & 0.3 & 0.3 & 0 & 0 \\ 0.3 & 1 & 0.3 & 0.3 & 0 & 0 \\ 0.3 & 0.3 & 1 & 0.3 & 0 & 0 \\ 0.3 & 0.3 & 0.3 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 \end{pmatrix}$$

As the variances of each variable are chosen to be 1, the covariance equals the correlation in this special case. The structure of Σ reveals that there is a block of four correlated variables and two uncorrelated ones.

– *Missing values*

Several MCAR, MAR and MNAR processes to create missing values were implemented. For each scheme, a given fraction $m \in \{0.0, 0.1, 0.2, 0.3\}$ of values is set missing for the variables X_2, X_4 and X_5. The number of observations that contain at least one missing value is given by $1 - (1 - \%_{\text{missing}})^{n_{\text{variables}}}$. Thus, a dataset that contains three variables with 30% missing values includes $1 - (1 - 0.3)^3 = 65.7\%$ incomplete observations on average. This seems to be a rather huge amount though it is not unlikely for real life data. Therefore, m comprises a wide range of possible scenarios.

In the MAR setting, the probability for missing values in a variable depended on the values of another variable. In the MNAR scheme this probability was determined by a variable's own values. Accordingly, each variable that contained missing values had to be linked to at least one other variable or itself. Table 2 lists the corresponding relations.

Table 2 List of variables that contain missing values or determine the probability of missing values

Contains missing values	Determines missing values	
(MCAR, MAR and MNAR)	(MAR)	(MNAR)
X_2	X_1	X_2
X_4	X_2	X_4
X_5	X_6	X_5

The schemes to produce missing values are:

– MCAR: Values are randomly set missing.
– MAR (rank): The probability of a value to be missing rises with the rank the same observation takes in the determining variable. More precisely, when m is the desired fraction of missing values and k is the index of the determining variable, then the probability of X_{ij} to be missing is: m times the rank of X_{ik} in X_k $(= r_{ik})$ divided by the sum of all ranks in X_k, i.e. $P(\text{“}X_{ij} \text{ is missing”}) = m \times r_{ik} / \sum_i r_{.k}$.
– MAR (median): The probability of a value to be missing is nine times higher for observations whose alue in the determining variable is above the corresponding median. Thus, $P(\text{“}X_{ij} \text{ is missing”}) = m \times 9/10$ when $X_{ik} > X_{[\text{med}]k}$ and $m \times 1/10$ else. [med] is the index of the median value.
– MAR (upper): Values of observations with the highest values in the determining variable are set missing.
– MAR (margins): Values of observations with the highest and lowest values in the determining variable are set missing.
– MNAR (upper): The highest values of a variable are set missing.

Independent test datasets served the purpose to evaluate the prediction error of a Random Forest. They were created in each simulation run exactly the same way as the training data. Therefore, they contained the same amount of missing values and were of equal size. The error was assessed by the MSE which equals the misclassification error rate in classification problems.

Random Forests consisted of 50 trees, each. The simulation was repeated 1,000 times. This high number of repetitions enabled a more precise estimation of summary statistics like the median importance

of a variable. More computational details are presented in Appendix B. Corresponding R-Code is given as online supplementary material.

In summary, there were two response types investigated for six processes to generate and three procedures to handle four different fractions of missing values. This sums up to as much as 144 simulation settings.

Simulation results

The following investigations are based on the classification analysis. Results for the regression problem are presented as supplementary material in Appendix A (Figures 5 and 6) as they showed similar properties.

The increased estimates for the importance of the correlated variables 1 and 2 compared to variable 5 are a general finding in each analysis, except for the null case. At first glance this might seem surprising as the association of these variables to the outcome was created using the same regression coefficients ($\beta = 1$). However, due to the correlation, it is correct to assign variables 1 and 2 a higher importance from a marginal point of view. The permutation importance measure follows this concept. As an alternative, Strobl et al. [53] introduced a conditional version of the permutation importance measure. It computes the importance of a variable conditioned on the values of other variables. It therefore tries to reflect conditional relations of variables to the outcome. Both concepts are justified and simply differ in the way to judge a variable's importance, i.e. in a marginal or conditional way. In some research fields the marginal perspective is used to investigate relations and interactions among variables [14, 54].

Another general finding is that there were no artificial effects observed for the non-influential variables in the null case (Figure 4 in Appendix A) or any other analysis setting (Figure 2).

Findings for the self-contained variable importance measure, which is able to implicitly deal with missing values, are displayed in Figure 2(a). According to expectations, the estimated importance of variables 2, 4 and 5 decreased as they contained a rising amount of missing values. It is interesting to note that meanwhile variable 1 is assigned a rising importance, although it does not seem to be directly affected. However, Hapfelmeier et al. [23] showed that this gain of relevance for prediction is justified: variables that are correlated and therefore provide similar information replace each other in a Random Forest when some of the information gets lost due to missing values. Accordingly, variable 1 takes over for variable 2 which is reflected in an increased selection frequency of variable 1 in the tree building process. In conclusion, this approach is allowed to be affected by the occurrence of missing values as it mirrors the situation at hand, i.e. the relevance a variable takes in a Random Forest under consideration of the information it actually provides. The self-contained importance measure appeared to be well suited for any of the missing data generating processes as results did not differ substantially.

Results for the complete case analysis, shown in Figure 2(b), showed undesired effects. A rising amount of missing values lead to a decreased estimate for the importance of the complete variable 1. This is partly due to the simple fact that some observations are completely discarded from analysis; importance measures typically return lower values when Random Forests are fit to less data. However, the estimate for variable 1 is not supposed to drop below that of variable 2 which contains the missing values. Unfortunately, this latter effect can be observed for every missing data generating process, except for MNAR (upper). It is most pronounced for MAR (upper) and MAR (margins). There is no rational justification for this property as variable 1 sustains its information while other variables lose it. A proper evaluation of a variable's relevance is supposed to reflect this fact. However, there is a simple explanation. For example, in the MNAR (upper) setting, the highest values of variable 1 cause some values in variable 2 to be missing. Now, as the corresponding observations are deleted from the data, variable 1 loses its highest values while variable 2 does not. This is why variable 1 loses more information then variable 2. As a consequence the estimated importance of variable 1 drops below that

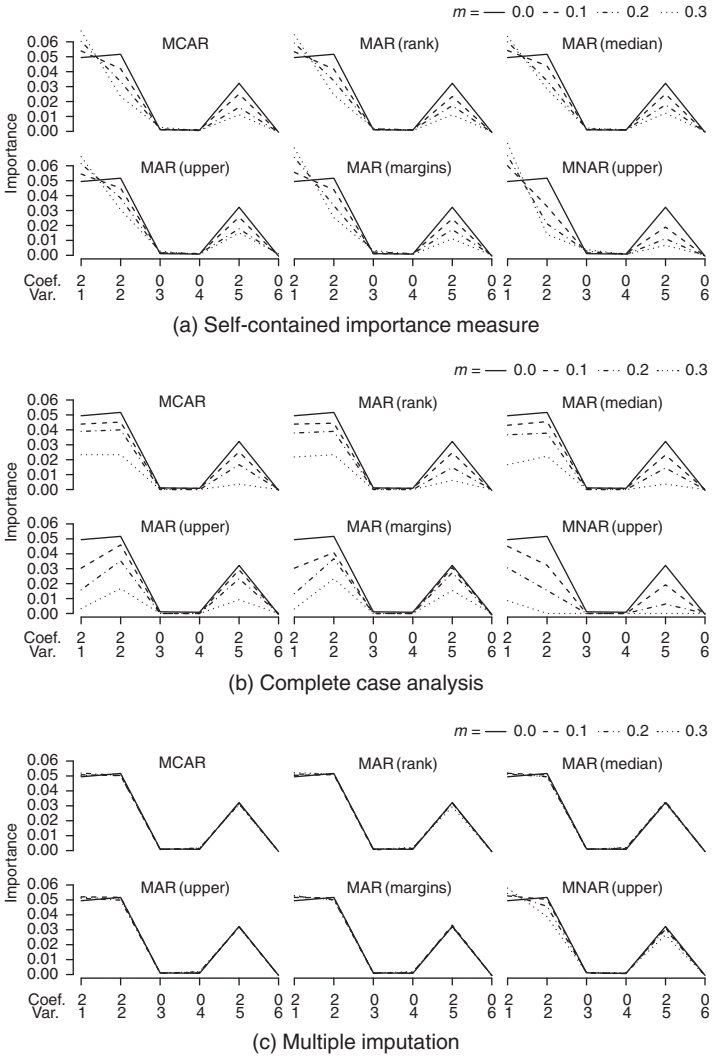


Figure 2 Median variable importance observed for the classification problem ($m = \%$ of missing values in X_2, X_4 and X_5)

of variable 2. In general, this effect is always present when important information of one variable causes missing values in another variable and in return is deleted by the complete case approach. That is why this effect can be seen, though less pronounced, in any of the MAR settings of 2b. Considering this vulnerability of complete case analysis to different missing data generating processes it should not be used for the assessment of a variable's importance when there are missing data.

An examination of Figure 2(c) reveals that multiple imputation, with only as few as five imputed datasets, is a convenient way to maintain and recover the importance of variables that would have been observed if there was no missing data at all. This equally held for variables that contained missing values and those which were completely observed; none of their estimated importances were considerably decreased or increased. Even the results for variable 5, which is only related to the outcome and therefore is associated with a rather weak imputation model remained unaffected by the amount of missing values. The example of variable 4 shows that imputation leads to artificially increased estimates for the importance of non-influential variables. However, this exceptionally good performance of the MICE algorithm can only be found in the MCAR and MAR settings. Here, the linear models used by the imputation algorithm are provided with all the information they need about the missing data generating processes. In addition, they are perfectly suited to deal with the normally distributed data and linear relationships. By contrast, in the MNAR setting, results are

closer to the ones produced by the self-contained importance measure designed to deal with missing values. This indicates that the MICE algorithm was not able to impute "proper" values in this data situation. As a consequence the variables with "badly" imputed values loose importance while their correlated competitors gain importance. In general, there are many reasons for a poor imputation that might lead to the same behavior even in MCAR or MAR settings, i.e. non-linearity.

The prediction error produced by each approach for the independent test sample is displayed in Figure 3. For multiple imputation it increases least with a rising amount of missing values. This effect is

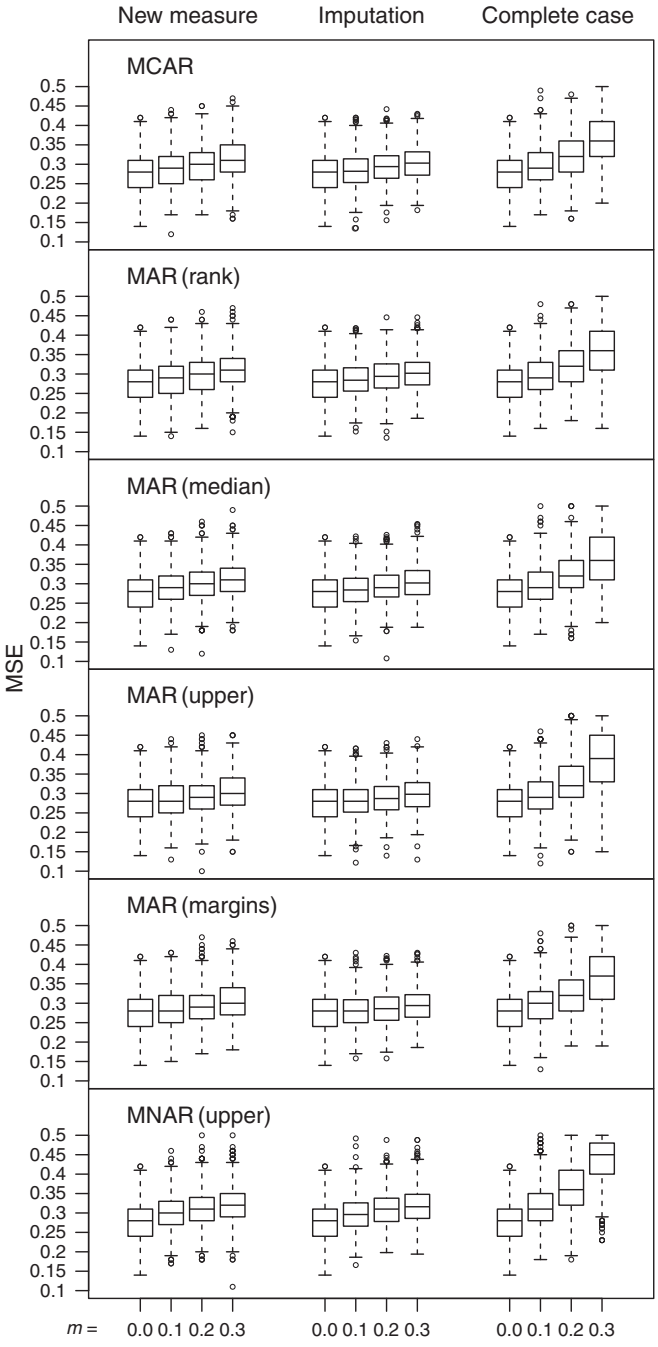


Figure 3 MSE observed for the classification problem (m = % of missing values in X_2, X_4 and X_5)

more pronounced for Random Forests that use surrogate splits; though there are only minor differences to multiple imputation (cf. Hapfelmeier et al. [10] and Rieger et al. [11], for a more detailed investigation of this issue). Complete case analysis appears to be much worse and leads to very high errors with a rising fraction of missing values. Missing data generating processes are comparable within each approach. However, there is one exception for the MNAR setting that always causes the worst $m = \%$ results. A corresponding evaluation of the regression problem is given as supplementary material in Appendix A (Figure 7).

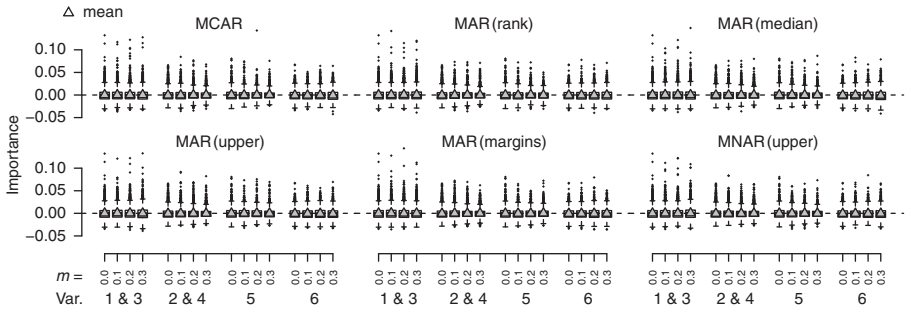
5 Conclusions

The analysis of liver surgery data showed that several laboratory blood tests, especially *lactate* and *bilirubin*, may be of interest for the prediction of postoperative complications. However, the estimation of their relevance as predictors varied between methods to handle missing values. Therefore, the properties of a self-contained importance measure designed to deal with missing values, complete case analysis and a multiple imputation approach have been investigated in additional simulation studies that employed several MCAR, MAR and MNAR processes to generate missing values. There are some clear recommendations for application: Unfavorable effects have been found for the complete case analysis in the MAR settings; the importance of completely observed variables was distorted in an undesirable way. This approach cannot be recommended for application. By contrast, the self-contained importance measure that implicitly deals with missing values expressed a decreased importance for variables that contained missing values. Therefore, it can be used to describe a variable's actual relevance for prediction in a Random Forest that was fit to data with missing values. This measure is especially useful for most of the everyday applications where Random Forests are usually fit to the raw data without any preprocessing of missing values. In some cases one might prefer to investigate the relevance a variable would have taken if there had been no missing values. Multiple imputation appeared to serve this purpose quite well when it is able to adequately replace the missing values. An additional evaluation of prediction error revealed that Random Forests that based on multiple imputed data were least affected by the occurrence of missing values. Results were only slightly worse when surrogate splits were used to process missing values. Complete case analysis lead to models with the highest prediction error. All of these results emphasize the significance of a deliberate decision about the use of an adequate analysis method.

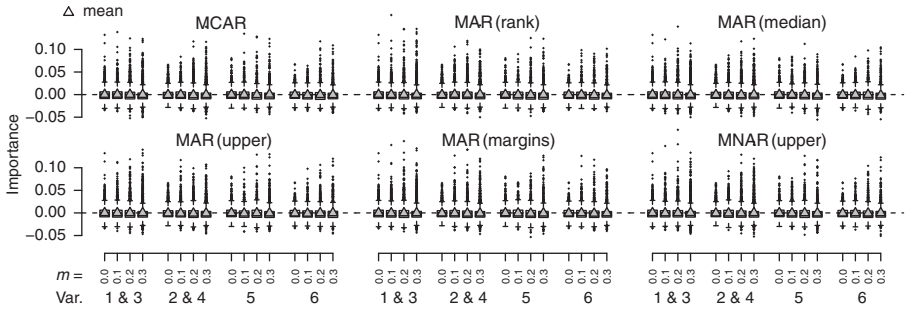
Appendix A

Supplementary material

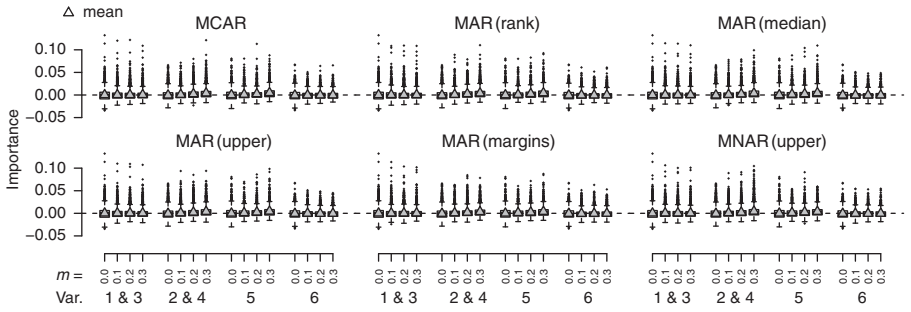
Figure 4 displays the estimated variable importances observed for the classification problem in the null case.



(a) Self-contained importance measure



(b) Complete case analysis



(c) Multiple imputation

Figure 4 Variable importance observed in the null case of the classification problem, $\beta = (0, 0, 0, 0, 0, 0)^\top$ ($m = \%$ of missing values in X_2, X_4 and X_5)

Figure 5 displays the estimated variable importances observed for the regression problem.

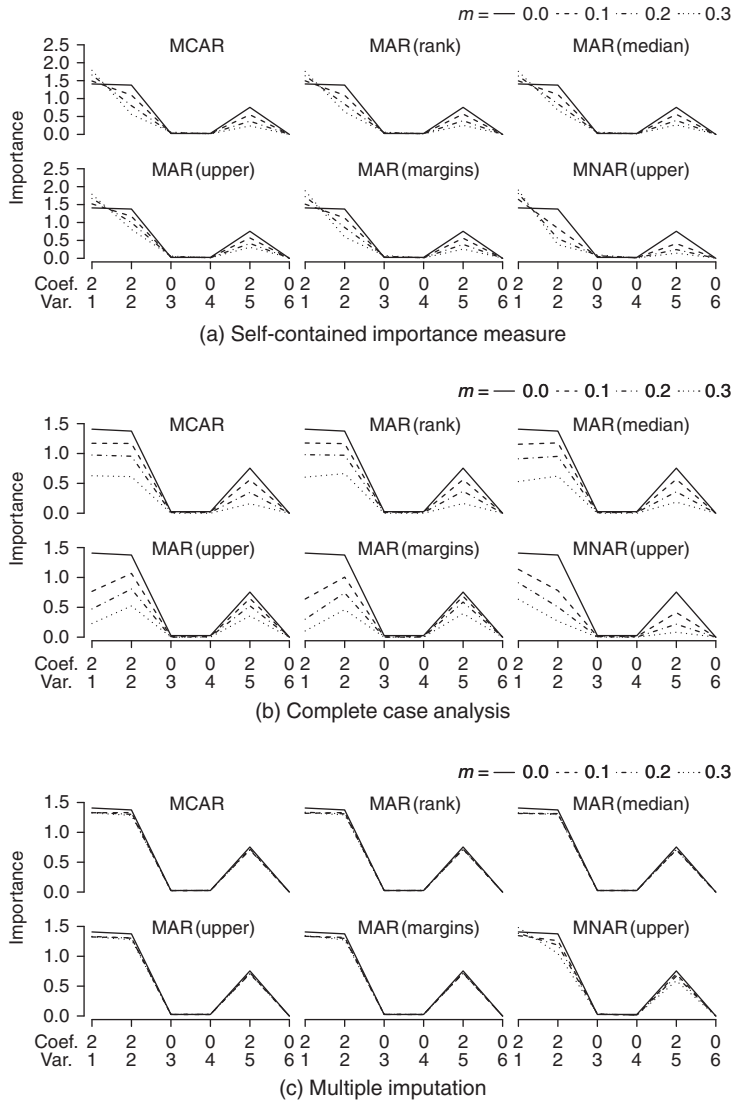


Figure 5 Median variable importance observed for the regression problem ($m = \%$ of missing values in X_2, X_4 and X_5)

Figure 6 displays the estimated variable importances observed for the regression problem in the null case.

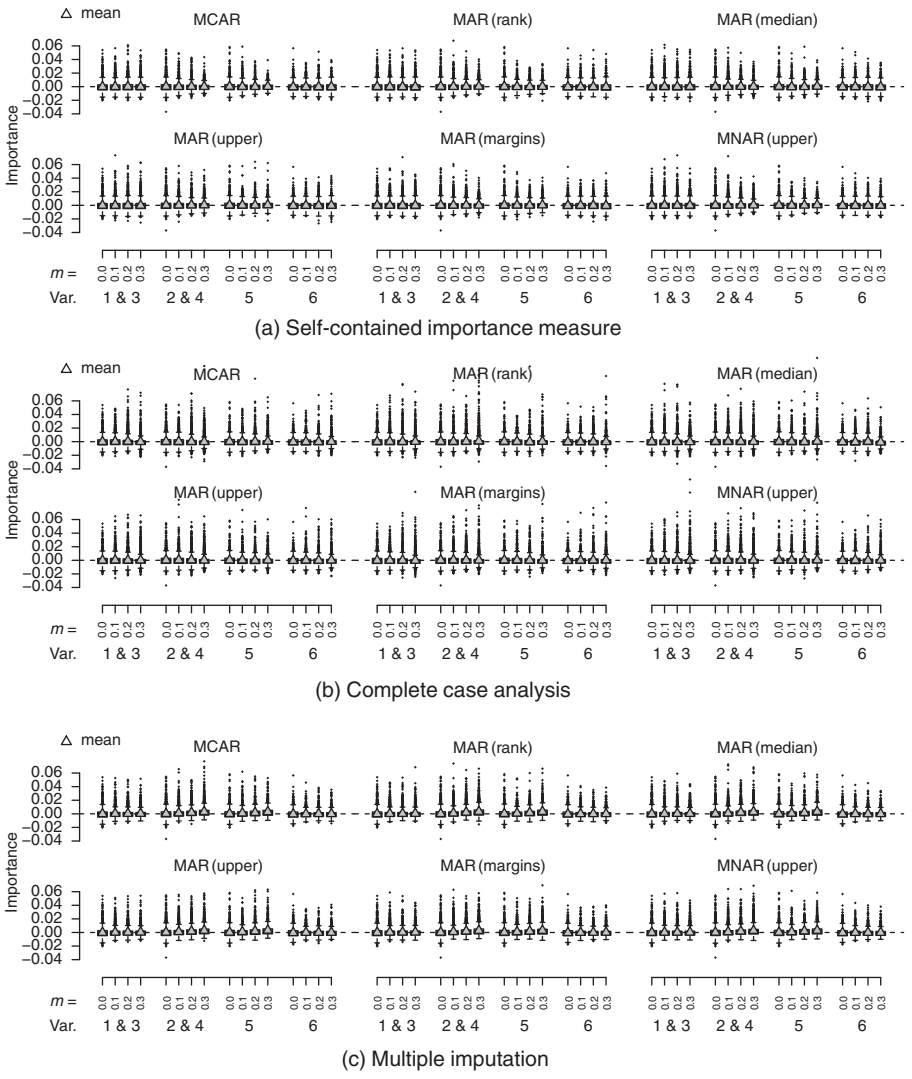


Figure 6 Variable importance observed in the null case of the regression problem, $\beta = (0, 0, 0, 0, 0, 0)^\top$ ($m = \%$ of missing values in X_2, X_4 and X_5)

Figure 7 displays the evaluation of prediction error for the regression problem.

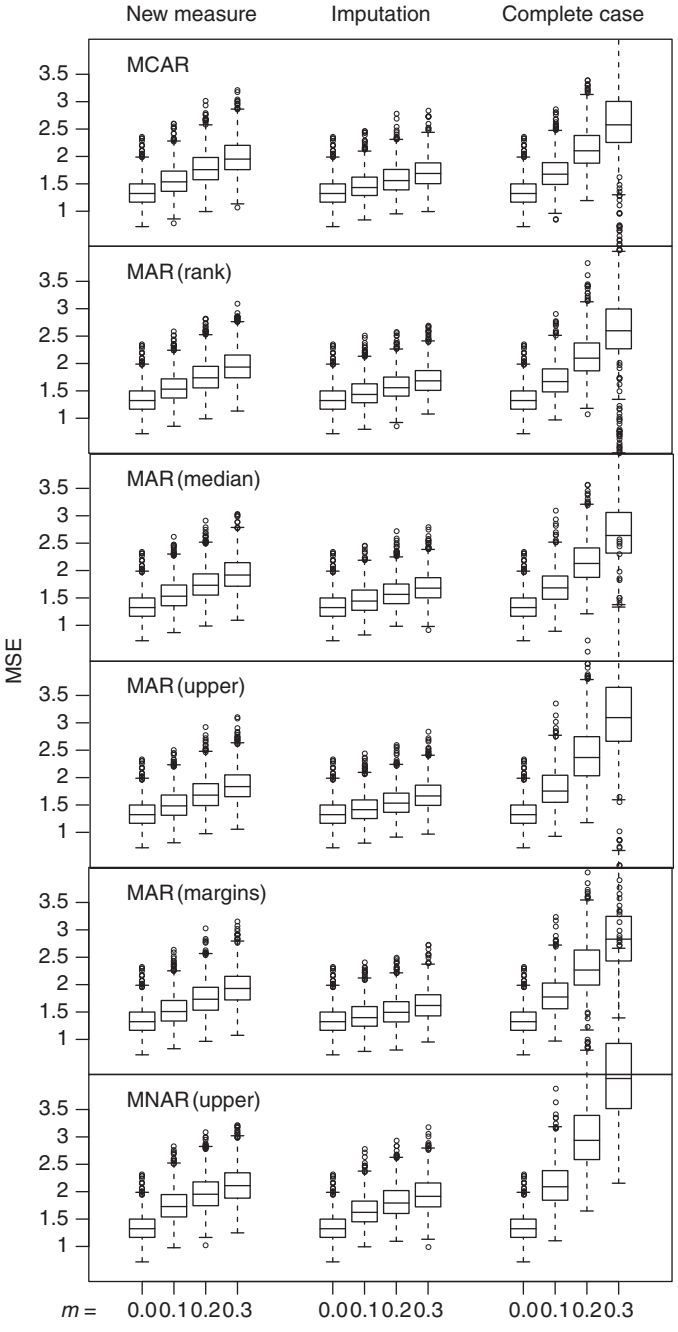


Figure 7 MSE observed for the regression problem ($m = \%$ of missing values in X_2, X_4 and X_5)

Appendix B

Computational details

The `R` system for statistical computing (version 3.0.1) [55] was used to implement the analyses. The package `party` (version 1.0.9) [56] provides unbiased Random Forests based on conditional inference by the function `cforest()`. Its settings were chosen to fit $ntree = 50$ trees in the repeated runs of the simulation study. For the analysis of the liver surgery data the number of trees was raised to $ntree = 500$. Each node was determined from $mtry = 3$ randomly selected variables and backed by $maxsurrogate = 3$ surrogate splits. There were no restrictions on the significance of a split ($mincriterion = 0$), and trees were grown until terminal nodes contained less than $minsplit = 20$ observations while child nodes had to contain at least $minbucket = 7$ observations. The computation of permutation importance measures was performed by the function `varimp()`. MICE is given by the function `mice()` of the package mice (version 2.18) [57]. It was used to produce five imputed datasets. A normal linear model was applied to impute continuous variables, a logistic regression for binary variables and a polytomous regression for variables with more than two categories; *defaultMethod* = c("norm", "logreg", "polyreg"). Each variable contributed to the imputation models. The fraction of imputed data in the simulation study was approximately $1 - (1 - m)^3$, $m \in \{0.0, 0.1, 0.2, 0.3\}$.

Appendix C

Investigation of data simulation

Details about the simulation study are given in Section 4. The data generating models made use of the parameter vector $\beta = (1, 1, 0, 0, 1, 0)^\top$. The estimation of these parameters by coefficients of linear and logistic regression models that were fit to the simulated data showed low bias and good coverage (Table 3).

Table 3 Bias and coverage of parameter estimates derived from the simulated data

Study	β	$\bar{\hat{\beta}}$	SE ($\hat{\beta}$)	Bias	Coverage (95%-CI)
Regression problem	1	1.000	0.080	0.000	0.944
	1	1.000	0.081	0.000	0.953
	0	−0.001	0.080	−0.001	0.945
	0	0.000	0.081	0.000	0.946
	1	0.999	0.074	−0.001	0.947
	0	0.003	0.073	0.003	0.957
	1	1.090	0.515	0.090	0.951
	1	1.127	0.522	0.127	0.943
Classification problem	0	0.010	0.442	0.010	0.951
	0	−0.002	0.456	−0.002	0.933
	1	1.118	0.506	0.118	0.927
	0	−0.006	0.399	−0.006	0.945

References

1. Aragon RJ, Solomon NL. Techniques of hepatic resection. J Gastrointest Oncol 2012;3:28–40.
2. Foster JH, Berman MM. Solid liver tumors. Major Problems in Clin Surg 1977;22:1–342.
3. Nagano Y, Togo S, Tanaka K, Masui H, Endo I, Sekido H, et al. Risk factors and management of bile leakage after hepatic resection. World J Surg 2003;27:695–8.
4. Ferlay J, Shin HR, Bray F, Forman D, Mathers C, Parkin DM. Estimates of worldwide burden of cancer in 2008: GLOBOCAN 2008. Int J Cancer 2010;127:2893–917.
5. Wu YL, Yu JX, Xu B. Safe major abdominal operations: hepatectomy, gastrectomy and pancreatoduodenectomy in elder patients. World J Gastroenterol 2004;10:1995–7.
6. Aloia TA, Fahy BN, Fischer CP, Jones SL, Duchini A, Galati J, et al. Predicting poor outcome following hepatectomy: analysis of 2313 hepatectomies in the NSQIP database. HPB 2009;11:510–15. Available at: http://dx.doi.org/10.1111/j.1477-2574.2009.00095.x.
7. de Meijer VE, Kalish BT, Puder M, Ijzermans JN. Systematic review and meta-analysis of steatosis as a risk factor in major hepatic resection. Br J Surg 2010;97:1331–9.
8. Lorenzo CS, Limm WM, Lurie F, Wong LL. Factors affecting outcome in liver resection. *HPB (Oxford)* 2005;7:226–30.
9. Breiman L. Random forests. Machine Learn 2001;45:5–32. Available at: http://dx.doi.org/10.1023/A:1010933404324.
10. Hapfelmeier A, Hothorn T, Ulm K. Recursive partitioning on incomplete data using surrogate decisions and multiple imputation. Comput Stat Data Anal 2012;56:1552–65. Available at: http://www.sciencedirect.com/science/article/pii/S0167947311003550.
11. Rieger A, Hothorn T, Strobl C. Random forests with missing values in the covariates, 2010. Available at: http://epub.ub.uni-muenchen.de/11481/.
12. Cutler DR, Edwards TC, Beard KH, Cutler A, Hess KT, Gibson J, et al. Random forests for classification in ecology. Ecology 2007;88:2783–92. Available at: http://www.esajournals.org/doi/abs/10.1890/07-0539.1.
13. Lunetta K, Hayward BL, Segal J, Van Eerdewegh P. Screening large-scale association study data: exploiting interactions using random forests. BMC Genet 2004;5:32. Available at: http://dx.doi.org/10.1186/1471-2156-5-32.
14. Altmann A, Tolosi L, Sander O, Lengauer T. Permutation importance: a corrected feature importance measure. Bioinformatics 2010;26:1340–7. Available at: http://dx.doi.org/10.1093/bioinformatics/btq134.
15. Archer K, Kimes R. Empirical characterization of random forest variable importance measures. Comput Stat Data Anal 2008;52:2249–60. Available at: http://dx.doi.org/10.1016/j.csda.2007.08.015.
16. Díaz-Uriarte R, Alvarez de Andrés S. Gene selection and classification of microarray data using random forest. BMC Bioinformatics 2006;7:3. Available at: http://www.biomedcentral.com/1471-2105/7/3.
17. Hapfelmeier A, Ulm K. A new variable selection approach using random forests. Comput Stat Data Anal 2013;60:50–69. Available at: http://www.sciencedirect.com/science/article/pii/S0167947312003490.
18. Hapfelmeier A, Ulm K. Variable selection with random forests for missing data, 2013. Available at: http://epub.ub.uni-muenchen.de/14344/.
19. Rodenburg W, Heidema AG, Boer JM, Bovee-Oudenhoven IM, Feskens EJ, Mariman EC, et al. A framework to identify physiological responses in microarray-based gene expression studies: selection and interpretation of biologically relevant genes. Physiol Genomic 2008;33:78–90. Available at: http://physiolgenomics.physiology.org/content/33/1/78.abstract.
20. Sandri M, Zuccolotto P. Variable selection using random forests. In Zani S, Cerioli A, Riani M, Vichi M, editors. Data analysis, classification and the forward search. Studies in classification, data analysis, and knowledge organization. Berlin, Heidelberg: Springer, 2006:263–70. Available at: http://dx.doi.org/10.1007/3-540-35978-8_30, 10.1007/3-540-35978-8_30.
21. Tang R, Sinnwell J, Li J, Rider D, de Andrade M, Biernacka J. Identification of genes and haplotypes that predict rheumatoid arthritis using random forests. BMC Proc 2009;3:S68. Available at: http://www.biomedcentral.com/1753-6561/3/S7/S68.
22. Yang W, Gu CC. Selection of important variables by statistical learning in genome-wide association analysis. BMC Proc 2009;3:S70. Available at: http://www.biomedcentral.com/1753-6561/3/S7/S70.
23. Hapfelmeier A, Hothorn T, Ulm K, Strobl C. A new variable importance measure for random forests with missing data. Stat Comput 2014;24:21–34. DOI: 10.1007/s11222-012-9349-1. Available at: http://dx.doi.org/10.1007/s11222-012-9349-1.
24. Horton NJ, Kleinman KP. Much ado about nothing: a comparison of missing data methods and software to fit incomplete data regression models. Am Statistician 2007;61:79–90. Available at: http://dx.doi.org/10.1198/000313007X172556.
25. Schafer JL, Graham JW. Missing data: our view of the state of the art. Psychol Methods 2002;7:147–77.
26. van Buuren S, Brand JP, Groothuis-Oudshoorn CG, Rubin DB. Fully conditional specification in multivariate imputation. Stat Comput Simulation 2006;76:1049–64.
27. White IR, Royston P, Wood AM. Multiple imputation using chained equations: issues and guidance for practice. Stat Med 2011;30:377–99. Available at: http://dx.doi.org/10.1002/sim.4067.
28. Janssen KJ, Donders AR, Harrell FE, Vergouwe Y, Chen Q, Grobbee DE, et al. Missing covariate data in medical research: to impute is better than to ignore. J Clin Epidemiol 2010;63:721–7. Available at: http://dx.doi.org/10.1016/j.jclinepi.2009.12.008.

29. Janssen KJ, Vergouwe Y, Donders AR, Harrell FE, Chen Q, Grobbee DE, et al. Dealing with missing predictor values when applying clinical prediction models. Clin Chem 2009;55:994–1001. Available at: http://dx.doi.org/10.1373/clinchem.2008.115345.
30. Rubin DB. Inference and missing data. Biometrika 1976;63:581–92. Available at: http://biomet.oxfordjournals.org/cgi/content/abstract/63/3/581.
31. Rubin DB. Multiple imputation for nonresponse in surveys. New York: J. Wiley & Sons, 1987.
32. Little RJ, Rubin DB. Statistical analysis with missing data, 2nd ed. Hoboken, New Jersey: Wiley-Interscience, 2002. Available at: http://www.worldcat.org/isbn/0471183865.
33. Strobl C, Boulesteix A-L, Augustin T. Unbiased split selection for classification trees based on the Gini index. Data Anal 2007;52:483–501. Available at: http://dx.doi.org/10.1016/j.csda.2006.12.030.
34. Harel O, Zhou X-H. Multiple imputation: review of theory, implementation and software. Stat Med 2007;26:3057–77.
35. Rubin DB. Multiple imputation after 18 + years. J Am Stat Assoc 1996;91:473–89. Available at: http://www.jstor.org/stable/2291635.
36. van Buuren S. Multiple imputation of discrete and continuous data by fully conditional specification. Stat Methods Med Res 2007;16:219–42. Available at: http://smm.sagepub.com/cgi/content/abstract/16/3/219.
37. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45:1–67. Available at: http://www.jstatsoft.org/v45/i03.
38. He Y, Zaslavsky AM, Landrum MB, Harrington DP, Catalano P. Multiple imputation in a large-scale complex survey: a practical guide. Stat Methods Med Res 2010;19:653–70. DOI: 10.1177/0962280208101273. Available at: http://smm.sagepub.com/cgi/content/abstract/0962280208101273v1.
39. Breiman L, Friedman J, Stone CJ, Olshen RA. Classification and regression trees, 1st ed. New York: Chapman & Hall/CRC, 1984. Available at: http://www.amazon.ca/exec/obidos/redirect?tag=citeulike09-20%&path=ASIN/0412048418.
40. Breiman L. Bagging predictors. Machine Learn 1996;24:123–40. Available at: http://dx.doi.org/10.1023/A:1018054314350.
41. Breiman L, Cutler A. Random forests, 2008. Available at: http://www.stat.berkeley.edu/users/breiman/RandomForests/cc_home.htm. Accessed 9 Jan 2013.
42. Dobra A, Gehrke J. Bias correction in classification tree construction. In Brodley CE, Danyluk AP, editors. Proceedings of the eighteenth international conference on machine learning (ICML 2001), Williams College, Williamstown, MA, USA. San Francisco, CA, USA: Morgan Kaufmann Publishers Inc. 2001:90–7.
43. Hothorn T, Hornik K, Zeileis A. Unbiased recursive partitioning. J Comput Graphical Stat 2006;15:651–74. Available at: http://pubs.amstat.org/doi/abs/10.1198/106186006X133933.
44. Kim H, Loh W. Classification trees with unbiased multiway splits. J Am Stat Assoc 2001;96:589–604.
45. White A, Liu W. Bias in information based measures in decision tree induction. Machine Learn 1994;15:321–9.
46. Strobl C, Boulesteix A-L, Zeileis A, Hothorn T. Bias in random forest variable importance measures: illustrations, sources and a solution. BMC Bioinformatics 2007;8:25. Available at: http://dx.doi.org/10.1186/1471-2105-8-25.
47. Ishwaran H. Variable importance in binary regression trees and forests. Electron J Stat 2007;1:519–37.
48. Venables WN, Ripley BD. Modern applied statistics with S, 4th ed. New York: Springer, 2002. Available at: http://www.stats.ox.ac.uk/pub/MASS4, ISBN 0-387-95457-0.
49. Austin PC, Tu JV. Bootstrap methods for developing predictive models. Am Statistician 2004;58:131–7. Available at: http://www.jstor.org/stable/27643521.
50. Sauerbrei W. The use of resampling methods to simplify regression models in medical statistics. J R Stat Soc Ser C (Appl Stat) 1999;48:313–29. Available at: http://dx.doi.org/10.2307/2680827.
51. Sauerbrei W, Royston P, Binder H. Selection of important variables and determination of functional form for continuous predictors in multivariable model building. Stat Med 2007;26:5512–28. Available at: http://dx.doi.org/10.1002/sim.3148.
52. Strobl C, Malley J, Tutz G. An introduction to recursive partitioning: rationale, application, and characteristics of classification and regression trees, bagging, and random forests. Psychol Methods 2009;14:323–48.
53. Strobl C, Boulesteix A-L, Kneib T, Augustin T, Zeileis A. Conditional variable importance for random forests. BMC Bioinformatics 2008;9:307. Available at: http://dx.doi.org/10.1186/1471-2105-9-307.
54. Nicodemus K, Malley J, Strobl C, Ziegler A. The behaviour of random forest permutation-based variable importance measures under predictor correlation. BMC Bioinformatics 2010;11:110. Available at: http://dx.doi.org/10.1186/1471-2105-11-110.
55. R Core Team. R: a language and environment for statistical computing. In: R Foundation for Statistical Computing, Vienna, Austria, 2013. Available at: http://www.R-project.org/.
56. Hothorn T, Hornik K, Strobl C, Zeileis A. Party: a laboratory for recursive part(y)itioning, 2008. Available at: http://CRAN.R-project.org/package=party, r package version 0.9-9993.
57. van Buuren S, Groothuis-Oudshoorn K. Mice: multivariate imputation by chained equations in R. J Stat Softw 2011;45:1–67. Available at: http://www.jstatsoft.org/v45/i03/.

Supplemental Material: The online version of this article (DOI: 10.1515/ijb-2013-0038) offers supplementary material, available to authorized users.